|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 학과 | 컴퓨터 전자시스템공학 | 학번 | 201904458 | 이름 | 이준용 |
| 구분 | 내용 | | | | |
| 학습 범위 | 기계학습 5장 딥러닝 최적화  5.3 규제의 필요성과 원리  5.4 규제 기법  5.5 하이퍼 매개변수 최적화  5.6 2차 미분을 이용한 최적화 | | | | |
| 학습 내용 | **기계학습 5장 딥 러닝 최적화**  **5.3 규제의 필요성과 원리**  과잉 적합에 빠지는 이유 -> 대부분 가지고 있는 데이터에 비해 훨씬 큰 용량의 모델을 사용하기 때문.  ➔ 훈련집합을 단순히 ‘암기’ 하는 형식이 되어버림.    현대 기계 학습은 충분히 큰 용량의 모델을 설계한 다음, 학습 과정에서 여러 규제 기법을 적용한다.  규제  모델 용량에 비해서 데이터가 부족한 경우의 불량 문제를 해결하기 위한 수단  적절한 가정을 투입하여 문제를 품  - 입력과 출력 사이의 매핑은 매끄럽다는 사전 지식.  -> ‘수치적 용량’을 제한하거나 비지도 학습을 이용.  - 오래 전부터 수학과 통계학에서 연구해온 주제.  Optical Flow : 인접한 Frame이 있을 때 Pixel의 움직임이 해당 개체를 어떻게 이동시키는가?  Nonlinear Image Registration : Deform을 시킬 때 해당 Field가 어떻게 변하느냐를 추적.    일반화 오류를 줄이려는 의도를 가지고 학습 알고리즘을 수정하는 방법 모두를 규제라고 할 수 있다.  **5.4 규제 기법**  \* 명시적 규제 : 가중치 감쇠나 드롭아웃처럼 목적함수나 신경망 구조를 직접 수정 방식  \* 암시적 규제 : 조기 멈춤, 데이터 증대, 잡음 추가, 앙상블처럼 간접적으로 영향을 끼치는 방식  가중치 벌칙  티호노프의 규제 기법 중 관련 변수가 드러나도록 할 경우에 해당  규제항은 훈련 집합과 무관하며 데이터 생성 과정에 내재한 사전 지식  매개변수를 작은 값으로 유지하게 하므로 모델의 용량을 제한 -> 수치적 용량 제한  MLP와 DMLP에 적용됨.  λR(θ)를 어떤 것으로, 얼마나 설정할 것인가? – 보통 L2놈이나 L1놈을 사용  \* 놈 : 벡터의 크기, 혹은 길이. L1 -> 절댓값의 합, L2 -> 자승의 합(유클리드 놈)    ∗ λ = 0으로 두면 규제를 적용하지 않은 원래 식 θ = θ - ρ∇J이 됨. -> 가중치 감쇠는 θ에 (1 − 2ρλ)를 곱해주기만  함.   * 최종 해를 원점 가까이 당기는 효과.   선형 회귀에 적용할 경우  공분산 행렬 𝑋𝑇𝑋의 대각 요소가 2λ만큼씩 증가  -> 역행렬을 곱하는 셈이 되므로 가중치를 축소하여 원점으로 당기는 효과  규제식 L1놈의 Gradient 계산    ➔ 부호만 고려한다   * L1놈의 희소성 효과 – 0이 되는 매개변수가 많음   -> 선형 회귀에 적용하면 특징을 선택하는 효과를 보여준다.  조기 멈춤  일정 시간(𝑇𝑎𝑝𝑡)이 지나면 과잉 적합 현상이 나타남 -> 일반화 능력이 저하 -> 훈련 데이터를 단순히 암기하는  문제가 발생  조기 멈춤은 검증집합의 오류가 최저인 점 𝑇𝑎𝑝𝑡에서 학습을 멈춘다.  실제 학습은 최저값 갱신이 한 번만 이루어 지지 않으므로 여러 최저값들의 parameter를 저장한 후 가장 낮은  최저값을 기록하는 parameter 𝑇𝑎𝑝𝑡를 선택하여 멈춘다.  데이터 확대  과잉적합을 방지하는 가장 확실한 방법으로, 큰 훈련집합을 사용한다.  -> 데이터 수집은 비용이 많이 든다는 한계가 있음.   * 데이터를 인위적으로 변형하여 확대함 : 자연계에서 벌어지는 잠재적인 변형을 프로그램으로 흉내   Ex : 모핑을 이용한 변형, 자연 현상 확대, 잡음 추가 등  드롭아웃  배치 정규화가 나오기 전까지 많이 사용된 방법으로, 입력층과 은닉층의 노드 중에서 임의로 선택하여 제거하는  방식   * 많은 부분 신경망을 자체적으로 제작하여 예측 단계에서 앙상블 결합 기법을 이용.   ➔ 계산 시간과 메모리 공간 측면 등, 비용에 대한 부담 문제가 발생.  앙상블  서로 다른 여러 개의 모델을 결합하여 일반화 오류를 줄인다.  서로 다른 예측기를 학습하는 일.  Ex : 배깅(훈련 집합을 여러 번 샘플링 하여 서로 다른 훈련집합을 구성), 부스팅(i번째 예측기의 오류를 i+1번째 에서 잘 인식하도록 의도적으로 구성)  이전에는 규제 기법으로 보지 않았으나 현대에 와서 규제에 대한 정의가 넓어짐에 따라 앙상블도 규제에 포함  **5.5 하이퍼 매개변수 최적화**  \* 내부 매개변수 : 에지 가중치로서 보통 θ로 표기, 학습 알고리즘이 최적화 함  \* 하이퍼 매개변수 : 모델의 외부에서 모델의 동작을 조정. Ex : 은닉층의 개수, CNN 마스크 크기와 보폭, 학습률 등.  하이퍼 매개변수 선택  표준 문헌이 제시하는 기본값을 사용   * 보통 여러 후보 값 또는 범위를 제시하며, 후보 값 중에서 주어진 데이터에 최적인 값을 선택함. * 하이퍼 매개변수 조합을 생성하는 방법에 따라 수동 탐색, 격자 탐색, 임의 탐색으로 나뉜다.   격자 탐색과 임의 탐색  ->보통 격자 탐색보다는 임의 탐색이 더 유리하다.  로그 규모 간격  어떤 매개변수는 로그의 규모를 사용해야 한다.  Ex : 학습률 볌위가 0.0001 ~ 0.1일 때   * 등간격 : 0.0001, 0.0002, 0.0003, …, 0.0998, 0.0999, 0.1 * 로그 규모 : 0.0001, 0.0002, 0.0004, 0.0008, …, 0.0256, 0.0512 …   매개변수가 m개이고 각각이 q개 구간이라면 𝑞  𝑚개의 점을 조사해야 함 -> 차원의 저주  **5.6 2차 미분을 이용한 방법**  \* 1차 미분을 이용한 방법 : 경사 하강법 – 현재 기계 학습의 주류 알고리즘  경사 하강법을 더 빠르게 할 수 있는가에 대해서 연구하다가 발견  뉴턴 방법  테일러 급수를 이용한 방법.  2차 함수에 뉴턴 방법을 적용했으므로 3차 항 이상을 무시한 식을 사용했음에도 최적의 경우를 제시함     * 기계 학습이 사용하는 목적 함수는 2차 함수보타 더 복잡한 형태이므로 한 번에 최적해에 도달하기는 힘   들다  ➔ 반복하는 뉴턴 방법을 사용해야 함.   * H를 구해야 함.   ➔ 매게 변수의 개수를 m이라 할 때 O(𝑚3  )라는 과다한 계산량 요구  ➔ 켤레 그레이디언트 방법이 대안으로 제시됨.  켤레 그레이디언트 방법  - 직선 탐색 방법  경사 하강법을 이용하여 얻은 해에 특정 값을 이용하여 해까지 일직선으로 도달하게 함  켤레 그레이디언트 방법의 식     * 𝑃𝑡와 𝑃𝑡−1를 켤레라고 부름.   유사 뉴턴 방법  H를 직접 구하는 대신 H의 역행렬을 근사하는 M을 사용한다.  처음에는 단위 행렬 I로 시작하여 그레이디언트 정보를 이용하여 점점 개선함.(LFGS)  기계 학습에서는 M을 저장하는 메모리를 적게 쓰는 L-BFGS를 주로 사용 | | | | |